PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-045268
(43)Date of publication of application : 16.02.1999

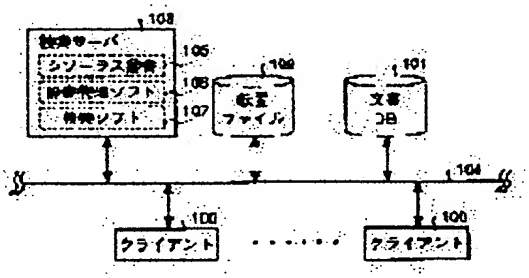
(51)Int.Cl. G06F 17/30
G06F 17/22
G06F 17/27
G06F 17/21

(21)Application number : 09-201985
(22)Date of filing : 28.07.1997

(71)Applicant : JUST SYST CORP
(72)Inventor : FUJITA SUMIO

(54) DOCUMENT RETRIEVAL DEVICE AND COMPUTER-READABLE RECORDING MEDIUM WHERE EPROGRAM MAKING COMPUTER FUNTION AS SAME DEVICE IS RECORDED

(57)Abstract:
PROBLEM TO BE SOLVED: To reduce the trouble of thesaurus dictionary management by registering noun phrases as relative words and similar words of corresponding index words in a relative word and similar word dictionary.
SOLUTION: Clients 100 select index words in the thesaurus dictionary 105 and specify the execution of a retrieving process for documents. Dictionary management softwar 106 manages the thesaurus dictionary 105. Retrieval software 107 generates a dislocation file 102 by using documents in a document data base 101, uses a retrieval engine to retrieve a corresponding document from the dislocation file 102 according to a selected index word in the thesaurus dictionary 105, and extracts an index word or similar word of the thesaurus dictionary 105 from the document as the retrieval result. Namely, a score corresponding statistical information on the appearance frequency, distribution, etc., of an object document group to be retrieved which is selected for the extracted noun phrase is imparted, the noun phrase having a score corresponding to a set retrieval condition is selected, and the noun phrase is registered as the relative word or synonym of the corresponding index word in the relative word or synonym.



LEGAL STATUS
[Date of request for examination]
[Date of sending the examiner's decision of rejection]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－45268

(43)公開日　平成11年(1999)２月16日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ	17/30	Ｇ０６Ｆ	15/40	３７０Ａ
	17/22		15/20	５２２Ｌ
	17/27			５５０Ｆ
	17/21			５７０Ｎ
			15/40	３７０Ｊ

審査請求　未請求　請求項の数４　ＯＬ　（全 11 頁）　最終頁に続く

(21)出願番号　特願平9－201985

(22)出願日　平成９年(1997)７月28日

(71)出願人　390024350
株式会社ジャストシステム
徳島県徳島市沖浜東３－46

(72)発明者　藤田　澄男
徳島市沖浜東３丁目46番地　株式会社ジャストシステム内

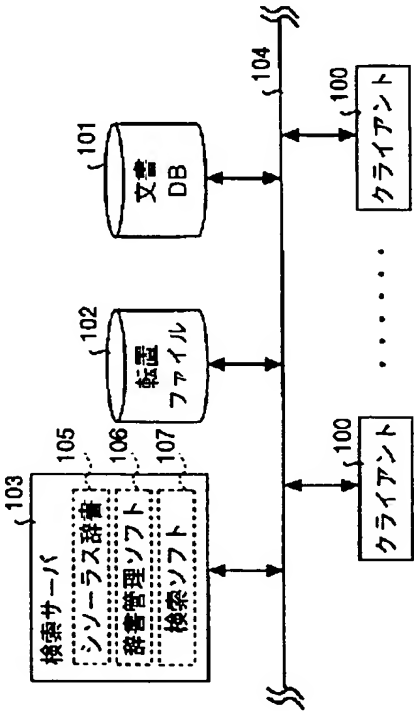
(74)代理人　弁理士　酒井　昭徳

(54)【発明の名称】　文書検索装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体

(57)【要約】

【課題】　検索によって得た文書から検索用のシソーラス辞書の索引語の関連語または類義語を自動的に抽出して登録できるようにすること。

【解決手段】　検索サーバ１０３は、意味体系に従って分類項目となる索引語を分類すると共に、関連語や類義語を該当する索引語にそれぞれ関連づけして記憶したシソーラス辞書１０５と、クライアント１００で索引語が選択されると、選択された索引語およびその関連語や類義語を検索条件として検索処理を行うと共に、検索結果の文書から名詞句を抽出し、抽出した名詞句に対し、選択された文書および検索対象の文書群における出現頻度および分布等の統計情報に応じたスコアを付与し、予め設定された選択条件に該当するスコアの名詞句を選択する検索ソフト１０７と、選択した名詞句を該当する索引語の関連語または類義語としてシソーラス辞書１０５に登録する辞書管理ソフト１０６と、を備えている。



【特許請求の範囲】
【請求項１】　検索条件に基づいて、検索対象の文書群から該当する文書を検索する文書検索装置において、
意味体系に従って分類項目となる索引語を分類すると共に、前記索引語の関連語および／または類義語を前記分類した索引語にそれぞれ関連づけして記憶した関連語・類義語辞書と、
前記関連語・類義語辞書の少なくとも索引語を画面表示する索引語表示手段と、
前記索引語表示手段で画面表示された索引語を選択するための索引語選択手段と、
前記索引語選択手段を介して索引語が選択されると、選択された索引語および前記索引語に関連づけられた関連語および／または類義語を前記検索条件として、該当する文書を検索する検索手段と、
前記検索手段で検索した文書の一覧を表示する一覧表示手段と、
前記一覧表示手段で表示された文書を選択するための文書選択手段と、
前記文書選択手段を介して選択された文書から名詞句を抽出する名詞句抽出手段と、
前記名詞句抽出手段で抽出した名詞句に対し、前記選択手段で選択した文書および検索対象の文書群における出現頻度および分布等の統計情報に応じたスコアを付与し、予め設定された選択条件に該当するスコアの名詞句を選択する名詞句選択手段と、
前記名詞句選択手段で選択した名詞句を該当する索引語の関連語または類義語として前記関連語・類義語辞書に登録する辞書登録手段と、
を備えたことを特徴とする文書検索装置。
【請求項２】　前記辞書登録手段は、登録しようとする名詞句が既に該当する索引語の関連語または類義語として前記関連語・類義語辞書中に存在する場合、前記存在する関連語または類義語に正の重みを付与し、
前記検索手段は、前記関連語または類義語に付与された重みを用いて、該当する文書の検索を行うことを特徴とする請求項１に記載の文書検索装置。
【請求項３】　前記文書選択手段は、前記一覧表示手段で一覧表示された文書を選択する際に、前記索引語に適合する文書に対して正の重み付けを指定を行うことが可能であると共に、前記索引語に適合しない文書に対して負の重み付けの指定を行うことが可能であり、
前記辞書登録手段は、前記正の重み付けが指定された文書から得た名詞句に正の重みを付与すると共に、前記負の重み付けが指定された文書から得た名詞句に負の重みを付与して該当する索引語の関連語または類義語として前記関連語・類義語辞書に登録し、
前記検索手段は、前記関連語または類義語に付与された重みを用いて、該当する文書の検索を行うことを特徴とする請求項１または２に記載の文書検索装置。
【請求項４】　前記請求項１～３のいずれか１つに記載の文書検索装置の各手段としてコンピュータを機能させるためのプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、検索によって得た文書から検索用のシソーラス辞書の索引語の関連語または類義語を自動的に抽出して登録できるようにした文書検索装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。
【０００２】
【従来の技術】複数の文書を格納した文書ＤＢ（データベース）から特定の文書を検索する文書検索装置は、一般に、検索式や検索文等の検索条件を入力し、入力した検索条件に該当する文書を文書ＤＢから検索するものである。
【０００３】ところで、上記文書検索装置では、入力した検索条件に基づいて検索を行うため、検索条件の語彙そのものではなく、検索条件中の語彙に関連する語彙を用いて記述された文書については、入力した検索条件に該当せず、検索結果に漏れが生じることがあった。
【０００４】そこで、検索用のシソーラス辞書を予め用意しておき、入力した検索条件を索引語として、該当する関連語や類義語をシソーラス辞書から抽出し、入力した検索条件にシソーラス辞書から抽出した関連語や類義語を加えて検索を行うことにより、検索結果に漏れが生じることを防止した文書検索装置が提案されている。
【０００５】
【発明が解決しようとする課題】しかしながら、上記従来の技術においては、検索用のシソーラス辞書を人手で生成しているため、常に最新の類義語が登録されている状態でシソーラス辞書を管理することは困難であるという問題点があった。特に、先端技術に関しては、常に新しい用語が次々と生まれてくるため、新たな用語を収集し、シソーラス辞書に登録する作業を継続的に行うことは困難であった。そして、シソーラス辞書への登録を怠れば、いくらシソーラス辞書を用いて検索を行ったとしても、常に高い精度の検索結果を得ることは不可能であるという問題点があった。
【０００６】本発明は上記に鑑みてなされたものであって、検索によって得た文書から検索用のシソーラス辞書の索引語の関連語または類義語を自動的に抽出して登録できるようにすることにより、シソーラス辞書を管理するための労力の軽減を図ることを目的とする。
【０００７】また、本発明は上記に鑑みてなされたものであって、シソーラス辞書を常に最新の関連語および類義語が登録された状態に保つことができるようにすることにより、精度の高い検索結果を得ることができるよう

にすることを目的とする。

【０００８】

【課題を解決するための手段】上記目的を達成するため、請求項１の文書検索装置は、検索条件に基づいて、検索対象の文書群から該当する文書を検索する文書検索装置において、意味体系に従って分類項目となる索引語を分類すると共に、前記索引語の関連語および／または類義語を前記分類した索引語にそれぞれ関連づけして記憶した関連語・類義語辞書と、前記関連語・類義語辞書の少なくとも索引語を画面表示する索引語表示手段と、前記索引語表示手段で画面表示された索引語を選択するための索引語選択手段と、前記索引語選択手段を介して索引語が選択されると、選択された索引語および前記索引語に関連づけられた関連語および／または類義語を前記検索条件として、該当する文書を検索する検索手段と、前記検索手段で検索した文書の一覧を表示する一覧表示手段と、前記一覧表示手段で表示された文書を選択するための文書選択手段と、前記文書選択手段を介して選択された文書から名詞句を抽出する名詞句抽出手段と、前記名詞句抽出手段で抽出した名詞句に対し、前記選択手段で選択した文書および検索対象の文書群における出現頻度および分布等の統計情報に応じたスコアを付与し、予め設定された選択条件に該当するスコアの名詞句を選択する名詞句選択手段と、前記名詞句選択手段で選択した名詞句を該当する索引語の関連語または類義語として前記関連語・類義語辞書に登録する辞書登録手段と、を備えたものである。

【０００９】また、請求項２の文書検索装置は、請求項１に記載の文書検索装置において、前記辞書登録手段が、登録しようとする名詞句が既に該当する索引語の関連語または類義語として前記関連語・類義語辞書中に存在する場合、前記存在する関連語または類義語に正の重みを付与し、前記検索手段が、前記関連語または類義語に付与された重みを用いて、該当する文書の検索を行うものである。

【００１０】また、請求項３の文書検索装置は、請求項１または２に記載の文書検索装置において、前記文書選択手段が、前記一覧表示手段で一覧表示された文書を選択する際に、前記索引語に適合する文書に対して正の重み付けを指定を行うことが可能であると共に、前記索引語に適合しない文書に対して負の重み付けの指定を行うことが可能であり、前記辞書登録手段が、前記正の重み付けが指定された文書から得た名詞句に正の重みを付与すると共に、前記負の重み付けが指定された文書から得た名詞句に負の重みを付与して該当する索引語の関連語または類義語として前記関連語・類義語辞書に登録し、前記検索手段が、前記関連語または類義語に付与された重みを用いて、該当する文書の検索を行うものである。

【００１１】さらに、請求項４のコンピュータ読み取り可能な記録媒体は、前記請求項１～３のいずれか１つに記載の文書検索装置の各手段としてコンピュータを機能させるためのプログラムを記録したものである。

【００１２】

【発明の実施の形態】以下、本発明の文書検索装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体の一実施の形態について、添付の図面を参照しつつ詳細に説明する。

【００１３】図１は、本実施の形態の文書検索装置のシステム構成図である。図１に示す文書検索装置は、後に詳細に説明するシソーラス辞書１０５中の索引語を選択することにより、文書の検索処理の実行を指定する機能を有する複数のクライアント１００と、シソーラス辞書１０５，シソーラス辞書１０５を管理する辞書管理ソフト１０６，および文書ＤＢ（データベース）１０１中の文書を用いて転置ファイル１０２を生成すると共に、ベクトル空間法を利用した検索エンジン（例えば、ＣＬＡＲＩＴＥＣＨ社のＣＬＡＲＩＴ等）を用いることにより、クライアント１００で選択されたシソーラス辞書１０５中の索引語に基づいて、転置ファイル１０２から該当する文書を検索し、さらに、検索結果の文書からシソーラス辞書１０５の索引語の関連語または類義語を抽出する検索ソフト１０７を備えた検索サーバ１０３と、上記クライアント１００や検索サーバ１０３等を接続するネットワーク１０４と、から構成されている。

【００１４】図１において、文書ＤＢ１０１は、クライアント１００等で作成された複数の文書を格納したものであり、格納される文書は、ワープロ文書や、ＳＧＭＬ，ＨＴＭＬ等の構造化文書等、いかなる種類の文書であっても良い。本実施の形態においては、文書ＤＢ１０１に格納された文書を検索対象とするが、検索対象を文書ＤＢ１０１中の文書に限定するものではない。

【００１５】転置ファイル１０２は、文書ＤＢ１０１中の複数の文書と、これら複数の文書それぞれから後述する方法で抽出した複数の索引語との関係を規定することにより、ある索引語が各文書それぞれにおいてどの程度重要な語彙であるかをベクター表現を用いて示したものであって、この索引語を用いて該当する文書を検索することができるようにしたものである。

【００１６】具体的には、１つの文書を複数のセンテンスからなるサブドキュメント単位に区切り、サブドキュメントから上記索引語となる名詞句を抽出して、抽出した名詞句それぞれについて、サブドキュメント中の出現頻度，文書ＤＢ１０１全体における分布等の統計情報を求め、求めた名詞句毎の統計情報を用いて各サブドキュメントをベクター表現に変換する。そして、変換したサブドキュメントのベクター表現に基づいて、文書のベクター表現を生成する。転置ファイル１０２は、このようにしてベクター表現された文書ＤＢ１０１中の文書を格納するものである。

【００１７】なお、各索引語には、対応する文書中の重要度に応じた重み付けを行うことができる。また、文書のベクター表現については、実際の検索を行う際に、サブドキュメントのベクター表現に基づいて生成することにしても良い。

【００１８】クライアント１００および検索サーバ１０３は、パーソナルコンピュータやワークステーション等によって構成される。

【００１９】図２は、検索サーバ１０３におけるシソーラス辞書１０５の内容を画面表示した様子の一例を示す説明図である。シソーラス辞書１０５は、意味体系に従って分類項目となる索引語を分類すると共に、索引語の関連語および／または類義語を分類した索引語にそれぞれ関連づけして記憶したものである。図２に示したシソーラス辞書１０５は、例として、新聞記事の記事分類に従い、階層構造となるように索引語を分類したものであり、フォルダ（またはディレクトリ）名のようにして表示されているものが索引語に該当する。また、ある索引語とその下位にある索引語とは、下位の索引語が上位の索引語の関連語または類義語となる。

【００２０】図２において、索引語「交通事故」について考えると、索引語「交通事故」は、索引語「社会面記事」の関連語となり（図２の左側の部分を参照）、下位の索引語「人身事故」，「物損事故」，「保険」は、それぞれ索引語「交通事故」の関連語または類義語となっている（図２の右側の部分を参照）。さらに、索引語「交通事故」には、図２の右側の部分に示すように、「衝突」，「死傷者」，「追突」，「脇見」，「飲酒」，「業務上過失致死」等が関連語または類義語として関連づけられている。なお、関連語および類義語には、索引語に対する関連性や類似性に応じて、それぞれ重みが付されており、付された重みを文書の検索の際に用いることができるようになっている。

【００２１】このシソーラス辞書１０５は、辞書管理ソフト１０６を介してクライアント１００からアクセスできるようになっており、図２は、クライアント１００でシソーラス辞書１０５の内容を画面表示した様子である。クライアント１００においては、検索したい文書が該当する索引語を探し、その索引語を図示しないマウス等で選択することにより、選択した索引語に関連づけられた関連語および／または類義語を用いた検索の実行を指定することができる。

【００２２】また、図３は、検索サーバ１０３において、検索ソフト１０７の処理を示す概略ブロック図である。検索ソフト１０７は、文書ＤＢ１０１中の文書を転置ファイル１０２に登録する処理と、ベクトル空間法を利用した検索処理と、検索結果の文書からシソーラス辞書１０５中の索引語の関連語または類義語を抽出する処理を行うものである。

【００２３】検索ソフト１０７において、転置ファイル１０２に登録する処理は、自然言語処理モジュール３００と、データベース・ビルド・コンポーネント３０４とによって行われる。

【００２４】具体的に、自然言語処理モジュール３００は、文書ＤＢ１０１から文書を入力し、文書のフォーマットの認識処理や、品詞情報等を格納した辞書３０１および各単語の係り受け等を解析するための文法辞書３０２を用いて形態素解析，構文解析，名詞句抽出等の解析処理を行い、上述したサブドキュメント毎の名詞句リストを含むドキュメント・セット３０３を生成する。

【００２５】データベース・ビルドコンポーネント３０４は、自然言語処理モジュール３００で生成したドキュメント・セット３０３を入力し、入力したドキュメント・セット３０３中の各サブドキュメントを上述したようにしてベクター表現に変換すると共に、サブドキュメントのベクター表現に基づいて、文書のベクター表現を生成して転置ファイル１０２に登録する。

【００２６】また、検索ソフト１０７において、文書の検索処理は、自然言語処理モジュール３００と、クエリー・ビルド・コンポーネント３０５と、検索エンジン３０７とによって行われる。

【００２７】具体的に、自然言語処理モジュール３００は、辞書管理ソフト１０６を介して、クライアント１００で選択されたシソーラス辞書１０５中の索引語とその索引語に関連づけられた関連語および／または類義語を検索条件として入力し、入力した検索条件の索引語・関連語または類義語毎にドキュメント・セット３０３を生成する。

【００２８】クエリー・ビルド・コンポーネント３０５は、ドキュメント・セット３０３を入力し、索引語，関連語または類義語について、文書ＤＢ１０１（転置ファイル１０２）全体における分布等の統計情報を求め、求めた統計情報と上記重みを用いてドキュメント・セット３０３をベクター表現に変換したクエリー・ドキュメント３０６を生成する。

【００２９】検索エンジン３０７は、クエリー・ビルド・コンポーネント３０５で生成したクエリー・ドキュメント３０６を入力し、転置ファイル１０２中の各文書のベクター表現とクエリー・ドキュメント３０６（索引語，関連語または類義語のベクター表現）とを比較して、クエリー・ドキュメント３０６との類似度に応じたスコアを各文書に付与し、所定の閾値を超えるスコアが付与された文書リスト３０８を検索結果として出力する。

【００３０】さらに、検索ソフト１０７において、シソーラス辞書１０５に登録する関連語または類義語の抽出処理は、自然言語処理モジュール３００と、シソーラス抽出エンジン３０９とによって行われる。

【００３１】具体的に、自然言語処理モジュール３００は、上述した検索処理の結果に基づいて、クライアント

１００で選択された文書を文書ＤＢ１０１から入力し、入力した文書について、フォーマットの認識処理や、品詞情報等を格納した辞書３０１および各単語の係り受け等を解析するための文法辞書３０２を用いて形態素解析，構文解析，名詞句抽出等の解析処理を行い、上述したサブドキュメント毎の名詞句リストを含むドキュメント・セット３０３を生成する。

【００３２】シソーラス抽出エンジン３１０は、自然言語処理モジュール３００で生成したドキュメント・セット３０３を入力し、入力したドキュメント・セット３０３中の各名詞句それぞれについて、各文書（ドキュメント・セット３０３）中の出現頻度や文書ＤＢ１０１（転置ファイル１０２）中の分布等の統計データを演算し、演算した統計データに基づいて、各名詞句にスコアを付与する。そして、予め設定した閾値を超えるスコアの名詞句を選択してシソーラスリスト３１０を生成し、辞書管理ソフト１０６に出力する。

【００３３】そして、辞書管理ソフト１０６は、検索ソフト１０３からシソーラスリスト３１０を入力し、シソーラスリスト３１０中の名詞句を、検索を開始する際にクライアント１００で選択された索引語の関連語または類義語としてシソーラス辞書１０５に登録する。

【００３４】なお、図１においては、文書ＤＢ１０１および転置ファイル１０２をネットワーク１０４に単独に接続した構成を示したが、これらを検索サーバ１０３に直接接続する構成としても良い。また、図１においては、本実施の形態の文書検索装置をネットワーク１０４を介したシステムで構成するように示したが、クライアント１００と検索サーバ１０３の処理を１つのコンピュータで行うようにすることもできる。

【００３５】次に、上述した構成を備えた文書検索装置の動作について、（１）転置ファイルの生成処理，（２）文書の検索処理，（３）シソーラス辞書への登録処理の順で詳細に説明する。

【００３６】（１）転置ファイルの生成処理

図４は、転置ファイルの生成処理を示すフローチャートである。検索サーバ１０３は、新たな文書が文書ＤＢ１０１に登録された場合（Ｓ４０１）、この文書を入力して転置ファイル１０２に登録するための処理を開始する（Ｓ４０２）。

【００３７】検索サーバ１０３において、自然言語処理モジュール３００は、ステップＳ４０２で入力した文書を解析する処理を行う（Ｓ４０３）。具体的には、入力した文書がワープロ文書，ＨＴＭＬ等の構造化文書等、いかなるフォーマットの文書であるかを判定する処理を行う。その後、辞書３０１および文法辞書３０２を用いて形態素解析，係り受け等の構文解析を行い、文書を複数のサブドキュメントに区分すると共に、区分したサブドキュメントから名詞句を抽出する等の処理を行う。

【００３８】そして、自然言語処理モジュール３００は、ステップＳ４０３における処理の結果に基づいて、サブドキュメント毎に名詞句リストを生成し、生成した名詞句リストを含むドキュメント・セット３０３を生成する（Ｓ４０４）。

【００３９】その後、データベース・ビルド・コンポーネント３０４は、自然言語処理モジュール３００で生成したドキュメント・セット３０３を入力し、文書のベクター表現を生成して転置ファイル１０２に登録する処理を行う（Ｓ４０５）。

【００４０】具体的には、ドキュメント・セット３０３中のサブドキュメントの各名詞句を転置ファイル１０２の索引語として、サブドキュメント中の出現頻度，文書ＤＢ１０１全体における分布等の統計情報を求め、求めた名詞句毎の統計情報を用いてサブドキュメントをベクター表現に変換する。この処理をドキュメント・セット３０３中の全てのサブドキュメントについて行い、変換したサブドキュメントのベクター表現に基づいて、文書のベクター表現を生成して転置ファイル１０２に登録する。その結果、文書ＤＢ１０１に新たに登録された文書が転置ファイル１０２に登録されることになる。

【００４１】（２）文書の検索処理

続いて、上述したようにして生成した転置ファイル１０２に基づいて、文書ＤＢ１０１中から特定の文書を検索するための処理について説明する。図５は、文書の検索処理を示すフローチャートである。

【００４２】ユーザは、クライアント１００を操作して、検索サーバ１０３の辞書管理ソフト１０７にシソーラス辞書１０５の内容の表示を要求する。その結果、クライアント１００に図２に示したシソーラス辞書１０５の内容が画面表示される。

【００４３】そこで、ユーザは、シソーラス辞書１０５中の索引語を参照し、検索によって得たい文書が該当する索引語を選択する。すなわち、この索引語は、検索条件の役割を果たすことになる。辞書管理ソフト１０７は、クライアント１００で索引語が選択されると（Ｓ５０１）、該当する関連語および／または類義語を検索条件として検索ソフト１０７に出力する（Ｓ５０２）。

【００４４】例えば、図２に示した「交通事故」という索引語がクライアント１００によって選択されたとすると、辞書管理ソフト１０６は、索引語「交通事故」と、索引語「交通事後」の関連語および／類義語を検索条件として検索ソフト１０７に出力することになる。

【００４５】検索ソフト１０７は、辞書管理ソフト１０６から検索条件を入力し、自然言語処理モジュール３００において、検索条件を構成する上記索引語，関連語および／または類義語毎にドキュメント・セット３０３を生成する（Ｓ５０３）。

【００４６】続いて、クエリー・ビルド・コンポーネント３０５は、自然言語処理モジュール３００からドキュメント・セット３０３を入力し、索引語，関連語または

類義語について、文書ＤＢ１０１（転置ファイル１０２）全体における分布等の統計情報を求め、求めた統計情報とそれらに付与された重みを用いてドキュメント・セット３０３をベクター表現に変換したクエリー・ドキュメント３０６を生成する（Ｓ５０４）。

【００４７】検索エンジン３０７は、クエリー・ビルド・コンポーネント３０５で生成したクエリー・ドキュメント３０６を入力し、転置ファイル１０２中の各文書のベクター表現とクエリー・ドキュメント３０６（索引語，関連語または類義語のベクター表現）とを比較して、クエリー・ドキュメント３０６との類似度に応じたスコアを各文書に付与する（Ｓ５０５）。すなわち、ベクトル空間法を用いた検索処理が行われる。

【００４８】なお、類似度に応じたスコアは、各文書とクエリー・ドキュメント３０６との類似度を余弦距離に基づいて表現したものであり、スコアが大きい文書がよりクエリー・ドキュメント３０６と類似していることを表している。

【００４９】そして、検索エンジン３０７は、予め設定されたスコアの閾値に基づいて、閾値を超えるスコアが付与された文書を選択し、選択した文書に基づいて、文書リスト３０８を生成し、クライアント１００に出力する（Ｓ５０６）。

【００５０】クライアント１００は、検索サーバ１０３から文書リスト３０８を入力し、入力した文書リスト３０８に基づいて、上位のランキングの文書から順に、該当する索引語に関連づけて文書一覧を表示する（Ｓ５０７）。クライアント１００のユーザは、一覧表示された文書から所望の文書を選択することにより、その文書を画面表示することができる。

【００５１】（３）シソーラス辞書への登録処理

さらに、検索結果の文書から名詞句を抽出してシソーラス辞書へ登録する処理について説明する。図６は、シソーラス辞書への登録処理を示すフローチャートである。

【００５２】クライアント１００のユーザは、画面表示された文書一覧から検索結果としてふさわしい文書（選択したシソーラス辞書１０５中の索引語に適合する文書）を選択し、選択した文書を検索結果として検索サーバ１０３に出力する。検索サーバ１０３の検索ソフト１０５は、クライアント１００から検索結果を入力すると（Ｓ６０１）、検索結果に該当する文書を文書ＤＢ１０１から入力する（Ｓ６０２）。

【００５３】文書ＤＢ１０１から文書を入力すると、自然言語処理モジュール２００は、入力した文書毎に、フォーマットの認識処理や、品詞情報等を格納した辞書２０１および各単語の係り受け等を解析するための文法辞書２０２を用いて形態素解析，構文解析，名詞句抽出等の解析処理を行う（Ｓ６０３）。

【００５４】その後、ステップＳ６０３における解析処理の結果に基づいて、サブドキュメント毎の名詞句リストを含むドキュメント・セット２０４を１文書を単位として生成する（Ｓ６０４）。

【００５５】シソーラス抽出エンジン３０９は、自然言語処理モジュール２００で生成したドキュメント・セット２０４を入力し、入力したドキュメント・セット２０４中の各名詞句それぞれについて、各文書（ドキュメント・セット２０４）中の出現頻度や文書ＤＢ１０１（転置ファイル１０２）中の分布等の統計データを演算する（Ｓ６０５）。

【００５６】ステップＳ６０５で統計データを演算した後、シソーラス抽出エンジン３０９は、求めた統計データに基づいて、各名詞句に対してスコア付けを行う（Ｓ６０６）。このスコアは、文書における各名詞句の重要性および検索を行う際に選択された索引語に対する関連性または類似性を表すもので、スコアが大きいもの程、重要性および関連性または類似性が高いことを表している。

【００５７】シソーラス抽出エンジン３０９は、ステップＳ６０６で行ったスコア付けの結果に基づいて、予め設定された閾値を超えるスコアの名詞句を、クライアント１００で選択された索引語（図５のステップＳ５０１参照）の関連語または類義語として抽出する（Ｓ６０７）。なお、ここでは、名詞句を抽出する条件として閾値を用いることにしたが、閾値に代えて、例えば、上位５番までのスコアの名詞句を抽出することにしても良い。

【００５８】その後、シソーラス抽出エンジン３０９は、ステップＳ６０７で抽出した名詞句のリストであるシソーラスリスト３１０を生成して、辞書管理ソフト１０６に出力する（Ｓ６０８）。

【００５９】辞書管理ソフト１０６は、検索ソフト１０７からシソーラスリスト３１０を入力し、入力したシソーラスリスト３１０中の名詞句を該当する索引語の関連語または類義語として、シソーラス辞書１０５に登録する（Ｓ６０９）。

【００６０】なお、辞書管理ソフト１０６は、登録しようとする名詞句が該当する索引語の関連語または類義語として既にシソーラス辞書１０５中に存在する場合には、存在する関連語または類義語に正の重みを付与する。したがって、検索を行う際に、これらの関連語または類義語を含む文書がヒットする率が高められる。

【００６１】また、検索された文書が一覧表示され、クライアント１００で一覧表示された文書から索引語に適合する文書を選択する際には、索引語に適合する文書に対して正の重み付けを指定を行うことができると共に、索引語に適合しない文書に対して負の重み付けの指定を行うことができる。そして、辞書管理ソフト１０６は、シソーラスリスト３１０の名詞句をシソーラス辞書１０５に登録する際に、正の重み付けが指定された文書から得た名詞句には正の重みを付与して関連語または類義語

として登録すると共に、負の重み付けが指定された文書から得た名詞句には負の重みを付与して関連語または類義語として登録する。その結果、正の重みが付与された関連語または類義語を含む文書が検索でヒットする確率が高まり、一方、負の重みが付与された関連語または類義語を含む文書が検索でヒットする確率は低められることになる。なお、検索結果の文書から得た名詞句のうち、正の重みの指定がなされた文書と負の重みの指定がなされた文書の両方に存在する名詞句については、シソーラス辞書１０５に登録しないようにすることもできる。

【００６２】さらに、図６のステップＳ６０９において、生成したシソーラスリスト３１０中の名詞句をそのままシソーラス辞書１０５に登録することにしたが、これらの名詞句を一度クライアント１００に画面表示し、ユーザが選択した名詞句のみをシソーラス辞書１０５に登録することにしても良い。

【００６３】このように、本実施の形態の文書検索装置によれば、検索によって得た文書からシソーラス辞書１０５中の索引語の関連語または類義語を自動的に抽出してシソーラス辞書１０５に登録できるようにすることにしたため、シソーラス辞書１０５を管理するための労力の軽減を図ることができると共に、シソーラス辞書１０５を常に最新の関連語や類義語が登録された状態に保つことができる。したがって、精度の高い検索処理を行うことができる。

【００６４】なお、本実施の形態においては、ベクトル空間法による検索を例にとって説明したが、ブーリアン検索により検索処理を行うことにしても良い。

【００６５】また、本実施の形態で説明した文書検索装置は、予め用意されたプログラムをコンピュータやワークステーションで実行することによって実現される。このプログラムは、ハードディスク，フロッピーディスク，ＣＤ－ＲＯＭ，ＭＯ，ＤＶＤ等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。また、このプログラムは、上記記録媒体を介して、またはネットワークを介して配布することができる。

【００６６】

【発明の効果】以上説明したように、本発明の文書検索装置（請求項１）によれば、意味体系に従って分類項目となる索引語を分類すると共に、索引語の関連語および／または類義語を分類した索引語にそれぞれ関連づけして記憶した関連語・類義語辞書と、関連語・類義語辞書の少なくとも索引語を画面表示する索引語表示手段と、索引語表示手段で画面表示された索引語を選択するための索引語選択手段と、索引語選択手段を介して索引語が選択されると、選択された索引語および索引語に関連づけられた関連語および／または類義語を検索条件として、該当する文書を検索する検索手段と、検索手段で検索した文書の一覧を表示する一覧表示手段と、一覧表示手段で表示された文書を選択するための文書選択手段と、文書選択手段を介して選択された文書から名詞句を抽出する名詞句抽出手段と、名詞句抽出手段で抽出した名詞句に対し、選択手段で選択した文書および検索対象の文書群における出現頻度および分布等の統計情報に応じたスコアを付与し、予め設定された選択条件に該当するスコアの名詞句を選択する名詞句選択手段と、名詞句選択手段で選択した名詞句を該当する索引語の関連語または類義語として関連語・類義語辞書に登録する辞書登録手段と、を備えたため、関連語・類義語辞書を管理するための労力の軽減を図ることができると共に、関連語・類義語辞書を常に最新の関連語や類義語が登録された状態に保つことができる。したがって、精度の高い検索処理を行うことができる。

【００６７】また、本発明の文書検索装置（請求項２）によれば、請求項１に記載の文書検索装置において、辞書登録手段は、登録しようとする名詞句が既に該当する索引語の関連語または類義語として関連語・類義語辞書中に存在する場合、存在する関連語または類義語に正の重みを付与し、検索手段は、関連語または類義語に付与された重みを用いて、該当する文書の検索を行うため、検索結果に応じて関連語・類義語辞書を学習させることができ、検索を重ねる毎に検索精度の向上を図ることができる。

【００６８】また、本発明の文書検索装置（請求項３）によれば、請求項１または２に記載の文書検索装置において、文書選択手段は、一覧表示手段で一覧表示された文書を選択する際に、索引語に適合する文書に対して正の重み付けを指定を行うことが可能であると共に、索引語に適合しない文書に対して負の重み付けの指定を行うことが可能であり、辞書登録手段は、正の重み付けが指定された文書から得た名詞句に正の重みを付与すると共に、負の重み付けが指定された文書から得た名詞句に負の重みを付与して該当する索引語の関連語または類義語として関連語・類義語辞書に登録し、検索手段は、関連語または類義語に付与された重みを用いて、該当する文書の検索を行うため、検索結果に応じて関連語・類義語辞書を学習させることができ、検索を重ねる毎に検索精度の向上を図ることができる。

【００６９】さらに、本発明のコンピュータ読み取り可能な記録媒体（請求項４）によれば、請求項１～３のいずれか１つに記載の文書検索装置の各手段としてコンピュータを機能させるためのプログラムを記録したため、このプログラムをコンピュータに実行させることにより、関連語・類義語辞書を管理するための労力の軽減を図ることができると共に、関連語・類義語辞書を常に最新の関連語や類義語が登録された状態に保つことができ、精度の高い検索処理を行うことができる文書検索装置を提供することができる。

【図面の簡単な説明】

【図１】本実施の形態の文書検索装置のシステム構成図である。

【図２】本実施の形態の文書検索装置において、検索サーバにおけるシソーラス辞書の内容を画面表示した様子の一例を示す説明図である。

【図３】本実施の形態の文書検索装置において、検索サーバにおける検索ソフトの処理を示す概略ブロック図である。

【図４】本実施の形態の文書検索装置において、転置ファイルの生成処理を示すフローチャートである。

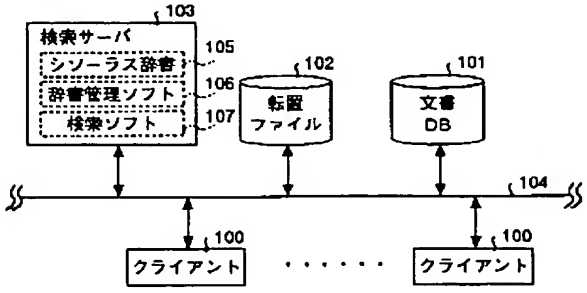
【図５】本実施の形態の文書検索装置において、文書の検索処理を示すフローチャートである。

【図６】本実施の形態の文書検索装置において、シソーラス辞書への登録処理を示すフローチャートである。

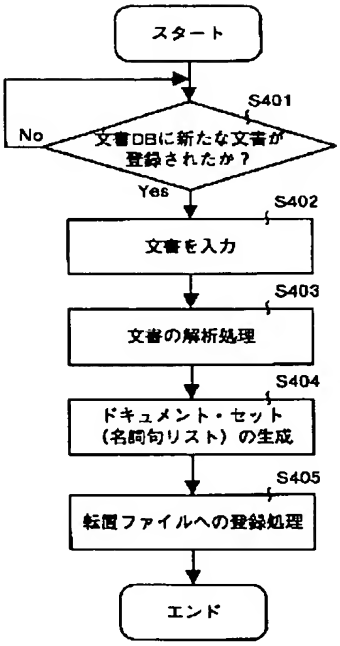
【符号の説明】

１００　クライアント
１０１　文書ＤＢ
１０２　転置ファイル
１０３　検索サーバ
１０４　ネットワーク
１０５　シソーラス辞書
１０６　辞書管理ソフト
１０７　検索ソフト
３００　自然言語処理モジュール
３０１　辞書
３０２　文法辞書
３０３　ドキュメント・セット
３０４　データベース・ビルド・コンポーネント
３０５　クエリー・ビルド・コンポーネント
３０６　クエリー・ドキュメント
３０７　検索エンジン
３０８　文書リスト
３０９　シソーラス抽出エンジン
３１０　シソーラスリスト

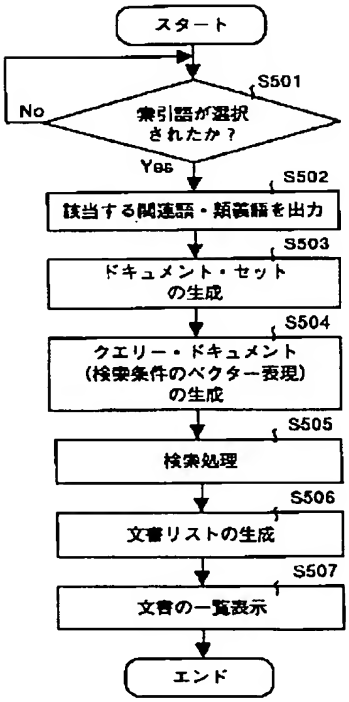
【図１】



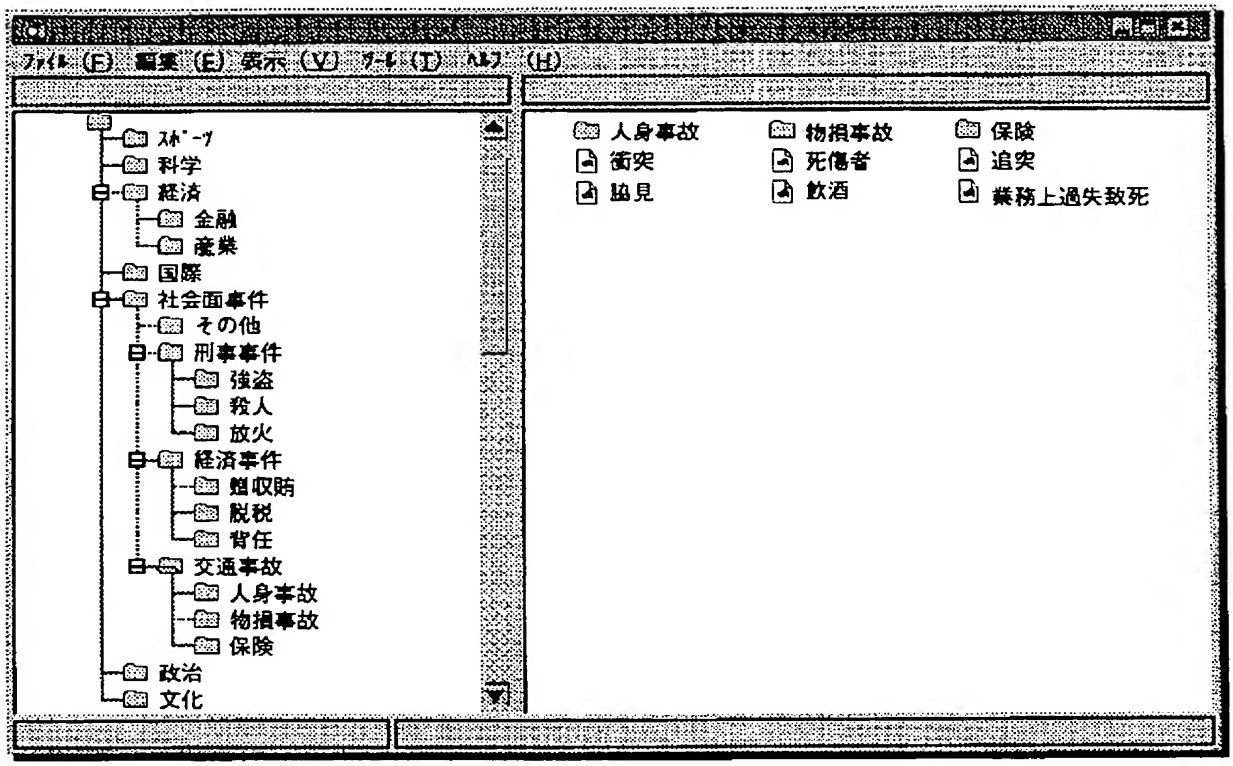
【図４】



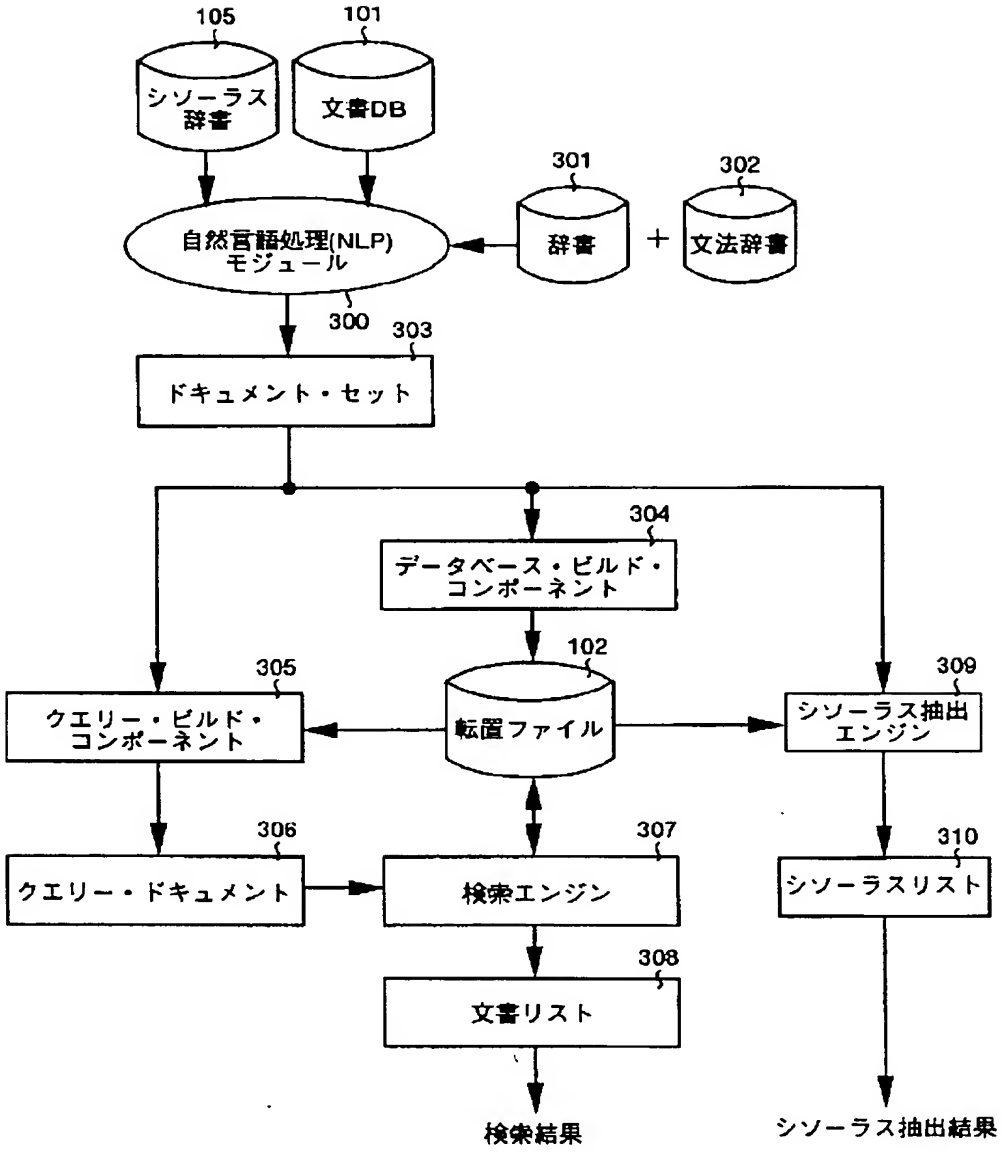
【図５】



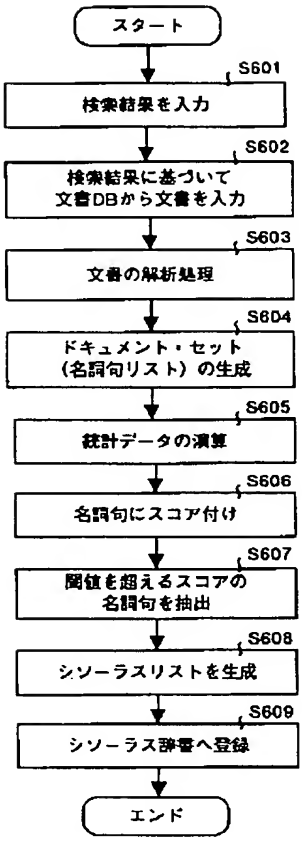
【図２】


ファイル(F)　編集(E)　表示(V)　ツール(T)　ヘルプ(H)
スポーツ
科学
経済
金融
産業
国際
社会面事件
その他
刑事事件
強盗
殺人
放火
経済事件
贈収賄
脱税
背任
交通事故
人身事故
物損事故
保険
政治
文化
人身事故
物損事故
保険
衝突
死傷者
追突
脇見
飲酒
業務上過失致死

【図３】


105
101
シソーラス辞書
文書DB
301
302
自然言語処理(NLP)モジュール
辞書
＋
文法辞書
300
303
ドキュメント・セット
304
データベース・ビルド・コンポーネント
102
305
309
クエリー・ビルド・コンポーネント
転置ファイル
シソーラス抽出エンジン
306
307
310
クエリー・ドキュメント
検索エンジン
シソーラスリスト
308
文書リスト
検索結果
シソーラス抽出結果

【図６】



フロントページの続き

(51) Int. Cl.[6]	識別記号	ＦＩ		
		Ｇ０６Ｆ	15/403	３２０Ｄ